DOMESTIC CANID VOCALIZATIONS: SITUATIONAL CONTEXT PREDICTION

A Thesis
Presented to
The Academic Faculty

by

Xiaochuang Han

In Partial Fulfillment
of the Requirements for the Degree
Bachelors of Science in the
College of Computing

Georgia Institute of Technology
April 2019

DOMESTIC CANID VOCALIZATIONS: SITUATIONAL CONTEXT PREDICTION

Approved by:

Dr. Melody Jackson, Advisor
School of Interactive Computing
Georgia Institute of Technology

Dr. Thad Starner
School of Interactive Computing
Georgia Institute of Technology

Date Approved: April 2019

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF SYMBOLS AND ABBREVIATIONS

FFT	Fast Fourier Transform
GMM	Gaussian Mixture Model
PCA	Principle Component Analysis
ICA	Independent Component Analysis
LDA	Latent Dirichlet Allocation
MFCC	Mel-Frequency Cepstral Coefficient
RMS	Root Mean Square

SUMMARY

Dogs live all around us and while we have little doubt they communicate with one another, our understanding of this communication is lacking. Our work aims to build a system that will be able to use artificial intelligence techniques to autonomously investigate the meaning behind a dog's vocalization. It will involve collecting data from a set of animals and recording both elicited and candid vocalizations. The data will then be separated into classes using machine learning as well as information gathered during data collection observation. Finally, the classes will be used to train an audio recognition model for real-time classification. This project's success could lead to new forms of rescue and service animal training as well as provide a basis on which other mammal vocalizations could be studied.

CHAPTER 1

INTRODUCTION

Research in canine communication has been stunted mainly because many ethologists considered domestic dogs to be so changed by artificial selection that their vocalizations lack any specific communicative functions. (Yeon, 2007) However, more recent research argues that not only are vocalizations by canines meaningful, the classifications are generalizable if the canine's physical size is taken into account. (Yin & McCowan, 2004) (Faragó, Pongrácz, Range, Virányi, & Miklósi, 2010) (Riede & Fitch, 1999) In addition, sophisticated frequency analysis of a mammalian vocalization has been developed. (Darden, Pedersen, & Dabelsteen, 2003) (Schrader & Hammerschmidt, 1997) While it is not likely that a Rosetta Stone of sophisticated canine language will be found, being able to reliably determine meaning behind vocalization will immensely improve our understanding of the animals that live the closest to us.

Canine vocalization studies have been carried out with frequency and statistical information interpreted by a human researcher. (Faragó, Pongrácz, Range, Virányi, & Miklósi, 2010) In contrast, this project aims to use machine learning techniques to automatically classify vocalizations and create a model that will be able to match newly heard vocalizations with pre-trained classifications. This approach will ideally provide insights into them that a human may have previously missed.

More specifically, we gathered data from dogs using an attached microphone and camera to capture the dog's vocalizations with high amplitude. We then extracted audio features and reduced the feature set using reduction algorithms. Finally, we run two sets of experiments. The first used the reduced dataset to train several classification

algorithms which were tested with a 20% test set and then 10 fold cross-validation. In the last experiment, a decision tree and logistic regression classifier was trained on the full feature set. This was done as these algorithms would provide feature importance which could also inform future feature picking.

The success of our work should produce a tool useful in the study of other mammalian vocalizations and have interesting effects on the training and deployment of service and support canines. Service animals could react to stimuli, such as a busy street intersection, as they naturally would then this system would quickly recognize the "danger" message and relay it verbally to the handler. Similarly, rescue animals can relay more specific information back to handlers using only their vocalization.

CHAPTER 2

METHODS AND MATERIALS

Our goal is to train a machine learning model that can detect, classify and interpret dogs' vocalization under different situational contexts. From a human's perspective, when we try to predict dogs' activity by their vocalizations, we first hear the sound made by the dog, then determine which kind of sound it is, such as bark or growl, and finally think about the type of possible canine activities based on our understanding of dogs' vocalization. Similarly, we constructed a three-step experiment for training the machine learning model using the same logic as from the human perspective.

Data Collection

A wired throat microphone connected to a GoPro camera was built for dogs to collect vocalization data. The owner of the dog positioned the microphone and camera at the dog's chest and left the lab. Two minutes later, the experimenter started to create one of the three situational contexts: "playing with human and another dog", "meeting a stranger", or "staying alone".

Pattern Detection

We set a hardcoded threshold for the vocalization audio's power level and its Fast Fourier Transform (FFT) spectrum's peak frequency. Through this process, we filtered out the background noise from our audio data and get a clean list of dogs' vocalizations. Then we used a Gaussian Mixture Model (GMM) to fit the FFT spectrums and Mel-frequency spectrograms of the vocalization. This operation helped us reduce the variance

in the audio data significantly. In the last, we run a series of component analysis algorithms on the mean frequency and power of the fitted Gaussians to separate different dimensions in different vocalizations. We will evaluate the resulting clusters and prove that they are consistent through different tests in the results chapter.

Target Prediction and Interpretation

Having recognized different vocalizations, we continued to predicting and reasoning the corresponding activities. In particular, we trained a decision tree and logistic regression model with the cluster info from the detection step as features and the activity type as labels. Then we looked for the top coefficients of the trained decision tree and logistic regression model which represents computer's "interpretation" of the relationship between dogs' vocalization and activity. We will further analyze and evaluate our model in the results chapter.

CHAPTER 3

RESULTS

Vocalization Clustering

We did 5 lab tests on 2 dogs in total and extracted 140 pure vocalization samples each with a length of 1 to 10 seconds from 3 classic scenarios: “playing with human and another dog”, “meeting a stranger”, and “staying alone”.

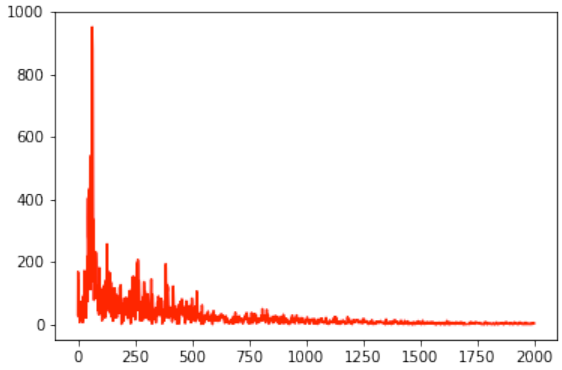

*Figure 1: 2-d FFT spectrum of a typical vocalization made in the **play** setting (x-axis: frequency in Hz, y-axis: power level, recording rate: 44100 samples per second, window size: 4096 samples, device: GoPro HERO with external microphone)*

	Gaussian mean	Gaussian weight
Low frequency Gaussian	181.64244267	0.68529098905
High frequency Gaussian	868.37391529	0.31470901095

Table 1: Gaussian mixture model fitted for Figure 1 (number of components: 2)

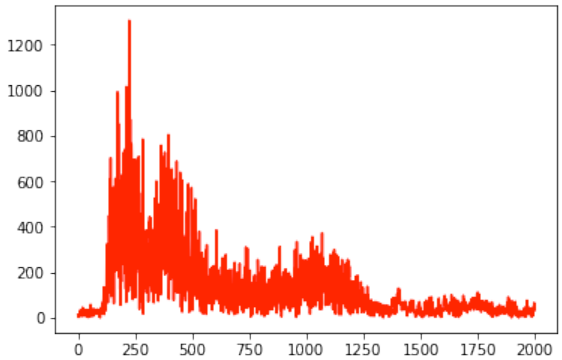

*Figure 2: 2-d FFT spectrum of a typical vocalization made in the **stranger** setting (x-axis: frequency in Hz, y-axis: power level, recording rate: 44100 samples per second, window size: 4096 samples, device: GoPro HERO with external microphone)*

	Gaussian mean	Gaussian weight
Low frequency Gaussian	322.14449396	0.523972193295
High frequency Gaussian	1046.83492694	0.476027806705

Table 2: Gaussian mixture model fitted for Figure 2 (number of components: 2)

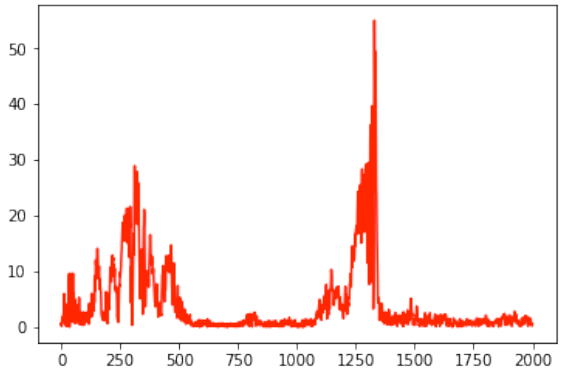

*Figure 3: 2-d FFT spectrum of a typical vocalization made in the **alone** setting (x-axis: frequency in Hz, y-axis: power level, recording rate: 44100 samples per second, window size: 4096 samples, device: GoPro HERO with external microphone)*

	Gaussian mean	Gaussian weight
Low frequency Gaussian	306.00760436	0.506550123863
High frequency Gaussian	1304.77725343	0.493449876137

Table 3: Gaussian mixture model fitted for Figure 3 (number of components: 2)

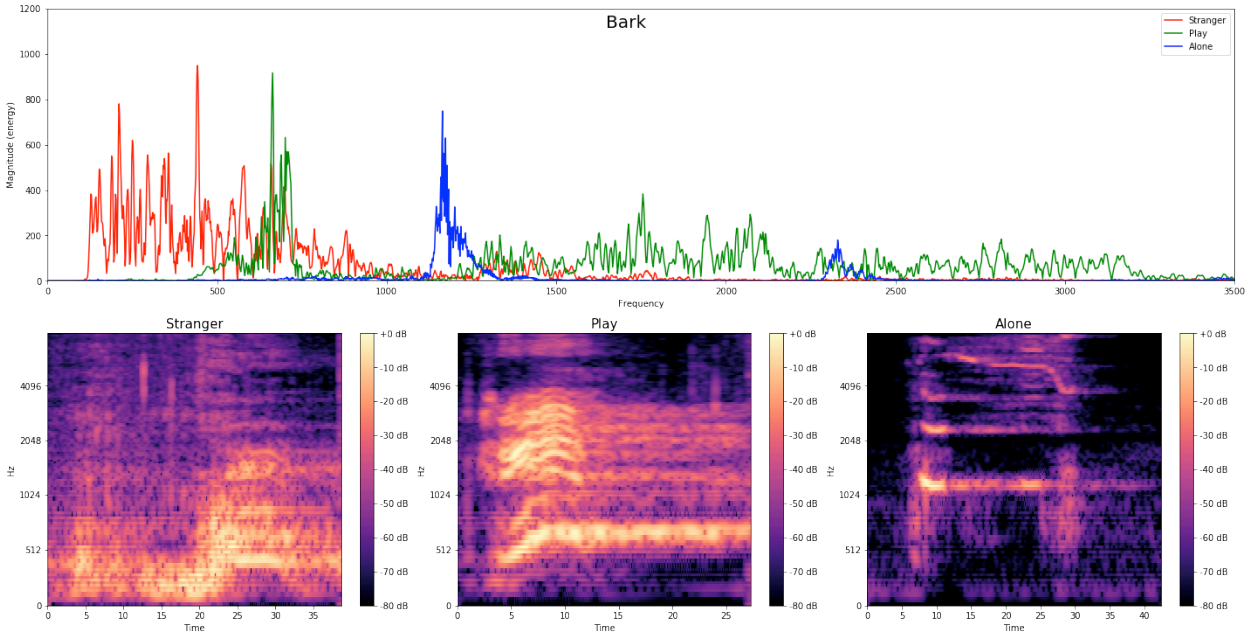

Figure 4: Mel-frequency spectrogram of typical barks made in the three settings (generated by a Python library, LibRosa)

As we can observe from the FFT spectrums and Mel-frequency spectrograms, the frequency and magnitude of the fitted Gaussians are differentiating indicators for predicting different situational contexts. For further verification that there exists a clear underlying clustering on the features set, we run four component analysis algorithms and found positive effect.

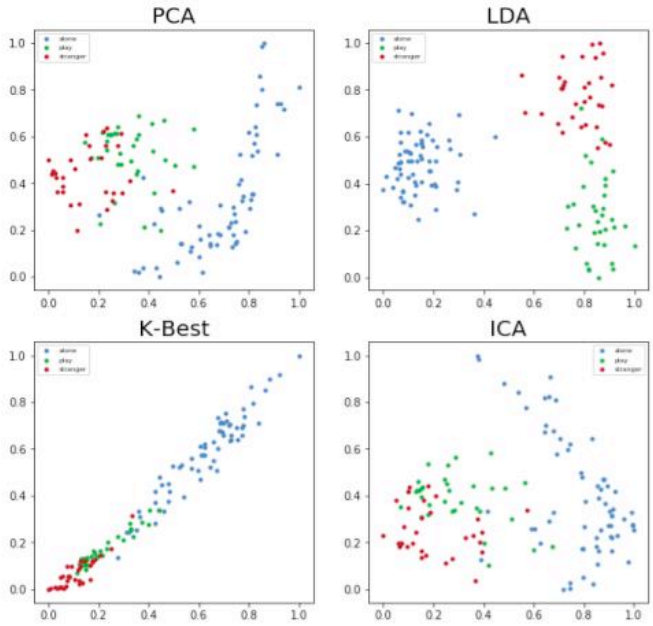

Figure 5: Four component analyses: PCA, LDA, K-Best, and ICA (green dots represent "play" setting samples, red dots represent "stranger" setting samples, and blue dots represent "alone setting samples)

Vocalization Classification

To train the decision tree and logistic regression classifier, we used the following features: *power level*, *logarithm of power level*, *mean of the low-frequency Gaussian peak*, *weight of the low-frequency Gaussian peak*, *mean of the high-frequency Gaussian peak*, *weight of the high-frequency Gaussian peak*, *zero crossing rate*, *MFCCs*, *wave bandwidth*, *wave centroid*, *wave flatness*, and *wave RMS*. Below are the best classifier's

accuracy and F1 score, decision tree's important features, and logistic regression's ablation study.

	Accuracy	F1
Decision Tree	0.857	0.86
Logistic Regression	0.834	0.818

Table 4: Accuracy and F1 score of the finely tuned classifiers

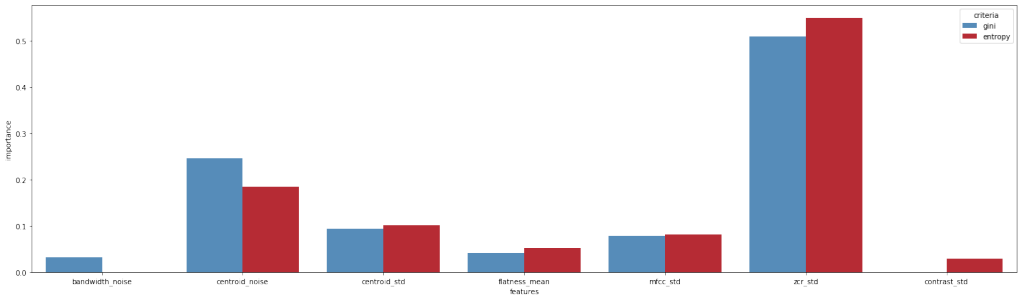

Figure 6: Feature importance of the Decision Tree classifier

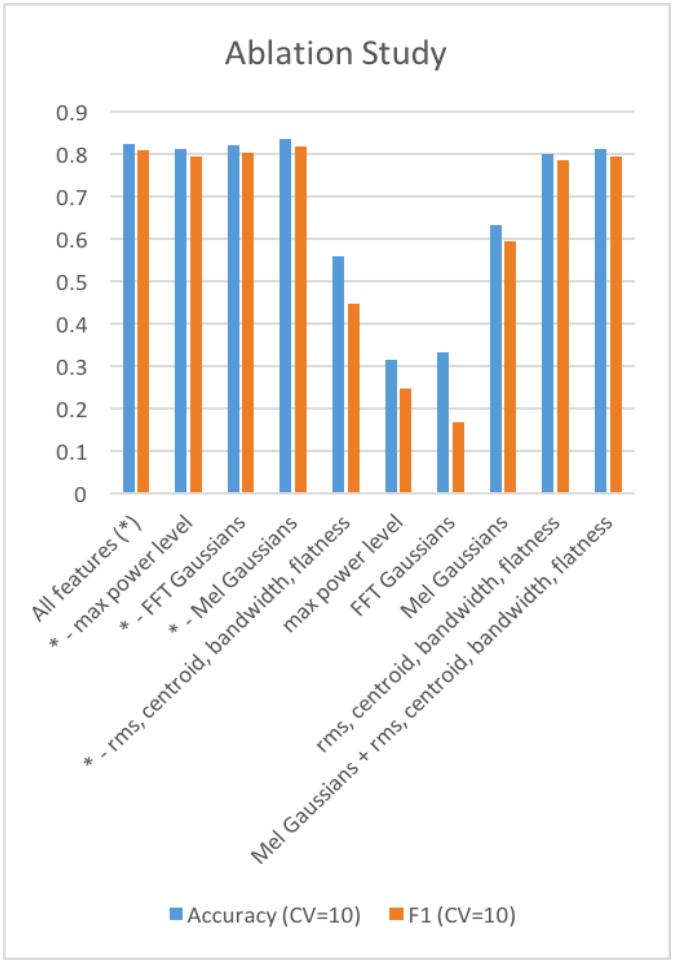

Figure 7: Ablation study of the Logistic Regression classifier

Our classifiers achieved a high accuracy and F1 score. The feature importance analysis and ablation study reflected that the Mel-frequency Gaussians, wave bandwidth, centroid, flatness, RMS, and zero crossing rate are crucial for the computer's prediction. We will discuss this more in the next chapter.

CHAPTER 4

DISCUSSION

We collected dog vocalizations in three classic settings of the research field: playing with human and another dog, meeting a stranger, and staying alone. Significant results were observed in FFT spectrums and Mel-frequency spectrograms of the typical vocalizations in the three settings. The vocalization made in the play setting has one clear peak in the low-frequency area (Figure 1). The vocalization made in the stranger setting also has one clear peak in the low-frequency area, but it also shows a broad and shallow peak in the high-frequency area (Figure 2). The vocalization made in the alone setting has a significantly low power level and it shows two rather even peaks in both low-frequency and high-frequency area (Figure 3).

Having different FFT spectrum and Mel-frequency patterns (Figure 4) for different vocalizations under different experiment settings corresponds well with the human interpretation of the vocalizations: playing bark, stranger growl, and alone whining. As a more quantitative result, the component analysis shows that there exists a clear underlying clustering over the features set (Figure 5).

As for the classification part, our best trained classifier achieved an accuracy of 85.7% and F1 score of 86.0%. The feature importance analysis and the ablation study show that the frequency patterns such as wave centroid and magnitude patterns such as zero crossing rate are crucial for the computer's prediction of situational contexts. Therefore, for future lightweight device building, we should focus on extracting these features prior to others.

REFERENCES

Darden, S. K., Pedersen, S. B., & Dabelsteen, T. (2003). METHODS OF FREQUENCY ANALYSIS OF A COMPLEX MAMMALIAN VOCALISATION. Bioacoustics-the International Journal of Animal Sound and Its Recording, 13(3), 247-263. Retrieved 1 26, 2018, from http://tandfonline.com/doi/abs/10.1080/09524622.2003.9753501

Faragó, T., Pongrácz, P., Range, F., Virányi, Z., & Miklósi, Á. (2010). 'The bone is mine': affective and referential aspects of dog growls. Animal Behaviour, 79(4), 917-925. Retrieved 1 25, 2018, from http://sciencedirect.com/science/article/pii/s0003347210000102

Riede, T., & Fitch, T. (1999). Vocal tract length and acoustics of vocalization in the domestic dog (Canis familiaris). The Journal of Experimental Biology, 202(20), 2859-2867. Retrieved 1 25, 2018, from https://ncbi.nlm.nih.gov/pubmed/10504322

Schrader, L., & Hammerschmidt, K. (1997). COMPUTER-AIDED ANALYSIS OF ACOUSTIC PARAMETERS IN ANIMAL VOCALISATIONS: A MULTI-PARAMETRIC APPROACH. Bioacoustics-the International Journal of Animal Sound and Its Recording, 7(4), 247-265. Retrieved 1 26, 2018, from http://tandfonline.com/doi/ref/10.1080/09524622.1997.9753338?scroll=top

Yeon, S. C. (2007). The vocal communication of canines. Journal of Veterinary Behavior-clinical Applications and Research, 2(4), 141-144. Retrieved 1 25, 2018, from http://journalvetbehavior.com/article/s1558-7878(07)00178-5/abstract

Yin, S., & McCowan, B. (2004). Barking in domestic dogs: context specificity and individual identification. Animal Behaviour, 68(2), 343-355. Retrieved 1 25, 2018, from http://sciencedirect.com/science/article/pii/s000334720400123x